HIDDEN MARKOV MODEL CLASSIFICATION OF MYOELECTRIC SIGNALS IN SPEECH

A. D. C. Chan[1,2], K. Englehart[1,2], B. Hudgins[1], D. F. Lovely[2]

[1]Institute of Biomedical Engineering, University of New Brunswick, Fredericton, Canada

[2]Department of Electrical and Computer Engineering, University of New Brunswick, Fredericton, Canada

Abstract – **A hidden Markov model based classifier is proposed in this paper to perform automatic speech recognition using myoelectric signals from the muscles of vocal articulation. The classifier's resilience to temporal variance is compared to a linear discriminant analysis classifier that was used in a pervious study. Speech recognition was performed, using five channels of myoelectric signals, on isolated words from a 10-word vocabulary. Temporal variance was induced by temporally misaligning data from the test set, with respect to the training set. When compared to the LDA classifier, the hidden Markov model classifier demonstrated a markedly lower variation in classification error due to the temporal misalignment. Characteristics of the hidden Markov model MES classifier suggest that it would effectively complement a conventional acoustic speech recognizer, in a multi-modal speech recognition system.**

Keywords – **automatic speech recognition, myoelectric signal, hidden Markov model**

I. INTRODUCTION

Automatic speech recognition (ASR) is a potential alternative control technology for high performance jet aircraft. ASR can significantly improve pilot efficiency and safety by simplifying the user interface and encouraging 'heads-up' flying [1]. Unfortunately, conditions during flight in a jet aircraft are not ideal for conventional ASR systems, which use only acoustic speech information to perform speech recognition. High ambient noise within the cockpit and various stress conditions that a pilot must endure while flying (e.g. high G-force, positive pressure breathing, vibration) degrade the classification accuracy of conventional ASR systems [1]. Work has been conducted to improve ASR under noisy conditions [2] and during speaker stress [3]; however, mono-modal ASR systems, relying solely on acoustic information, will eventually saturate in performance.

Recently, we proposed a multi-modal ASR system, using the myoelectric signal (MES) from articulatory muscles as a second source of speech information [4]. MES has two advantages in ASR. First, MES is immune to acoustic noise. Second, word pairs that sound similar but are articulated in a dissimilar manner, manifest distinctively in the MES (e.g. "sign" and "fine"). While a conventional ASR would have difficulty distinguishing between these words because of their acoustic similarity, the distinctiveness in the MES enables the MES ASR to differentiate these signals with relative ease.

In a previous study investigating this multi-modal approach, speech information was shown to be present in the MES by performing ASR, using only MES from five articulatory muscles [4]. A linear discriminant analysis (LDA) classifier was utilized on a set of wavelet transform features, reduced by principle component analysis (PCA). Classification errors ranged from 2.68% to 10.36% for a 10-word vocabulary.

While this result was encouraging, the LDA classifier required temporal alignment of the data, which was accomplished by aligning the MES data from each word repetition to the start of the audible speech. Temporal variance was further reduced by instructing subjects to maintain a constant speaking rate. In practical situations, speaking rate will vary, which poses a problem for MES ASR. The temporal position of articulation relative to the acoustic signal will vary with speaking rate; therefore so will the relative position of the MES data. If the MES data are temporally misaligned, the classification error for the LDA classifier will increase.

In this paper, the use of a hidden Markov model (HMM) classifier is proposed to perform ASR on the MES. Most current conventional ASR systems use HMMs to classify acoustic speech information. HMMs use a Markov chain topology, which preserves the structural characteristics and temporal ordering of the signal. In addition, each state in the Markov chain has statistical parameters, which account for the probabilistic nature of the observed data. The structure of the HMM allows it to cope with time-scale variance and shape variance in the observed signal. Thus, the HMM classifier is expected to be much more resilient to temporal variance than the LDA classifier.

II. METHOD AND MATERIALS

In this study, ASR was performed on isolated words from a 10-word vocabulary, using the information in the MES. The resilience of the HMM to temporal variations was evaluated by classifying words from the test set that were temporally misaligned with the training set. The performance of the HMM classifier was compared to the LDA classifier used in the previous study [4].

The MES data set used in this study was a subset of the data used in the previous study [4]. Surface MES were obtained from five articulatory muscles of the face: the *levator anguli oris*, the *zygomaticus major*, the *platysma*, the *depressor anguli oris*, and *the anterior belly of the digastric*. Each MES channel was collected using pairs of Ag-AgCl button electrodes, embedded in the lining of a fighter pilot's oxygen mask (Fig. 1). Electrodes were 1/2" in

0-7803-7211-5/01$10.00©2001 IEEE

Report Documentation Page

Report Date 15 Oct 2001	**Report Type** N/A	**Dates Covered (from... to)** -

Title and Subtitle Hidden Markov Model Classification of Myoelectric Signals in Speech	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Institute of Bioemdical Engineering University of New Brunswick Fredericton, Canada	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

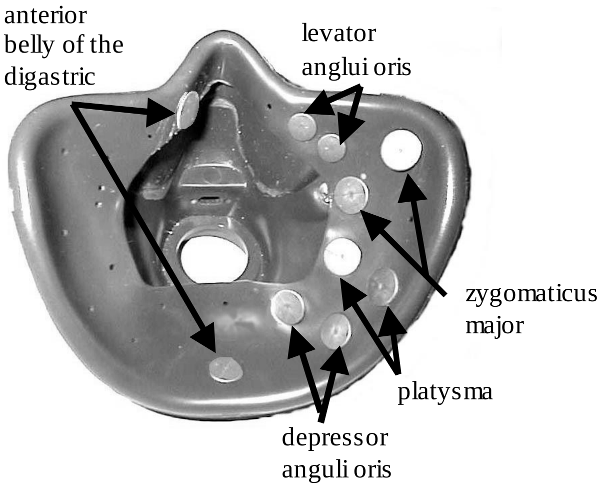

Fig. 1. Pilot oxygen mask with electrodes embedded in the lining

diameter, except for the *levator anguli oris* electrodes, which were 3/8" in diameter. A light coating of electrode paste was applied to each electrode to improve the electrode-skin interface. A Red-Dot^{TM} Ag-AgCl electrode was placed on the back of the neck to provide a common reference. The five MES channels were differentially amplified and simultaneously sampled at 1000 Hz, along with the acoustic speech signal, which was obtained using a cardiod dynamic microphone, held by the subject near the opening at the front of the mask.

Two Canadian English-speaking male subjects (S1 and S2) participated in this study. Both subjects had no known speech disorders. A 10-word vocabulary, consisting of the words "zero" to "nine" was used. For each subject, 15 series of 30 words were constructed, each series containing three repetitions of each word in the vocabulary[1]. The order of the words in each series were randomly permuted and presented to the subject one at time, with at least one second between words to minimize coarticulatory and anticipatory effects. Subjects were asked speak each word in a consistent manner, minimizing the variation in volume and speaking rate. A rest period between each series of words was provided for subjects, as needed.

The data were processed offline. The data set from each subject was split in half, with every second word used as part of the training set and the remaining words used as the test set. MES data for each word repetition were segmented into records of 1024 ms, using the acoustic channel as a trigger. In the previous study, it was found that there exists speech information in the MES up to 500 ms preceding the start of the audible speech [4]; therefore, the training set was segmented using a fixed pretrigger value of 500 ms. To misalign the test set with respect to the training set, a pretrigger value ranging from 400 to 600 ms, in steps of 25 ms, was used to segment the test set. This resulted in a ±100 ms range of temporal misalignment of the test set with respect to the training set.

ASR was performed using two classification techniques. The first classification technique was the LDA classifier used in the previous study [4]. Wavelet transform coefficients were computed for each MES channel and these coefficients underwent PCA feature reduction. Six PCA coefficients were retained from each MES channel, presenting a total of 30 features to the LDA classifier.

The second classification technique used a six state, left-right HMM, with single mixture observation Gaussian densities. Observation windows of 64 ms were used. In each observation window, three features were extracted from each MES channel: the first two autoregressive coefficients, and the integrated absolute value [5]. It was found, empirically, that using higher order autoregressive coefficients did not improve the classification rate for this setup. For each observation window, an observation vector (o_i) of dimension 15 was computed (3 features × 5 MES channels). Observation windows overlapped with a spacing of 8 ms between windows, so the observation sequence ($O = [o_1, o_2, ..., o_N]$) had a length of $N = 121$. A HMM (λ_j) was trained on the observation sequence for every word in the vocabulary (W_j; $j = 1, 2, ..., 10$), using the expectation-maximization algorithm [6]. An unknown word (W_k) from the test set was classified by first computing its observation sequence (O_k). Next, the likelihood of each HMM generating that observation sequence was computed ($\mathrm{P}(O_k|\lambda_j)$; j = 1, 2, ..., 10). Finally, the unknown word was classified according to the maximum likelihood (i.e. $W_k = W_\ell$, where $\ell = \arg\max_j \mathrm{P}(O_k|\lambda_j)$).

III. Results

Fig. 2 is a plot of the classification error as a function of temporal misalignment for both classifiers and both subjects. Positive temporal misalignment corresponds to a decrease in the pretrigger value used in data segmentation of the test set.

At a temporal misalignment of 0 ms, where the test set is properly aligned with the training set, the LDA classifier has classification errors of 10.36% (S1) and 2.68% (S2), while the HMM classifier has classification errors of 13.06% (S1) and 14.73% (S2). As the temporal misalignment increases in the positive or negative direction, the classification error of the LDA classifier increases significantly for both subjects. The maximum classification error for the LDA classifier, within the ±100 ms range of temporal misalignment, is 55.41% (S1) and 44.64% (S2), an increase of over 40% from the minimum classification error. Temporal misalignment did not have any considerable

[1] The total number of words was intended to be 450; however, six words for subject S1 and one word for subject S2 were mistakenly not recorded (clipped at the beginning or end of the data collection).

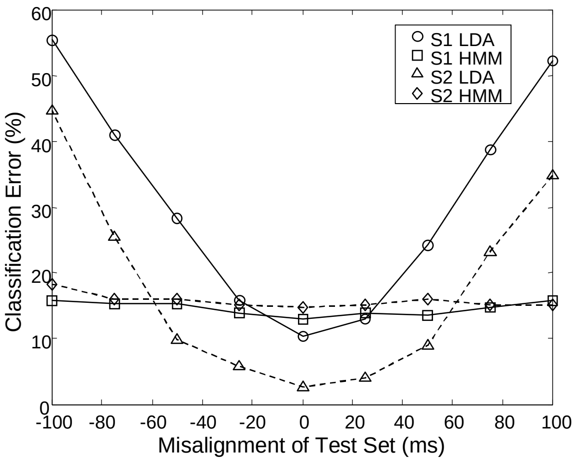

Fig. 2. Classification error as a function of temporal misalignment.

effects on the classification rate of the HMM classifier. The maximum classification error of the HMM classifier reaches 15.76% (S1) and 18.27% (S2), which differs from its minimum classification error by only 2.70% (S1) and 3.57% (S2).

Two additional observations were made for the HMM classifier, which have important implications for a multi-modal ASR system. First, the confusion matrices for the HMM classifier (Fig. 3 and Fig. 4) show that two words can account for approximately half of the total misclassifications for each subject. For subject S1, 12 of the 29 misclassifications resulted from the words "zero" and "six". For subject S2, 20 of the 33 misclassifications resulted from the words "eight" and "nine".

Second, during misclassification, typically there was only a small difference between the likelihoods computed for the correct word, and the likelihood computed for the incorrectly chosen word. With such a small difference in likelihoods, perhaps we should not be classifying just to a single word. Consider, instead, using the MES ASR system simply to reduce the dimensionality of the classification problem for a second stage of a multi-modal system. In the first stage, the MES ASR system chooses the two words with the highest likelihoods (i.e. classification on the second rank). Now, the second stage would only have to distinguish between those two words. Classifying the data on the second rank, the classification errors of the MES ASR system are 5.86% and 6.70%, for subject S1 and S2, respectively. Thus, the MES ASR reduced the 10-word classification problem to a 2-word classification problem with high accuracy.

IV. Discussion

A HMM classifier was presented in this paper to perform ASR on MES. As expected, the HMM classifier had a markedly lower variation in classification error due to temporal misalignment, compared to the LDA classifier. This is because the HMM uses a time sequence of observations to classify its data, instead of the examining the data in its entirety, which is how the LDA classifier operates.

Although the HMM classifier has demonstrated a superior resilience to temporal variance, when there is little or no temporal misalignment, the LDA classifier has a lower classification error than the HMM classifier. However, the HMM classification error may be improved by optimizing the signal features used in the observation vectors. Also, the current structure of the HMM classifier does not account for the dependence or correlation between observations, which can be accomplished by including state duration parameters and time derivatives of features; this may further improve the classification rate.

The HMM-based MES ASR system has already demonstrated that it is able to reduce a 10-word classification problem to a 2-word classification problem, with an accuracy exceeding 93%. In addition, two words in

		HMM Word Classification									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Word Spoken	**0**	17		1			1	2	1		
	1		20		1					1	
	2			20				2			
	3	1			20		1				
	4					20	2				
	5				1		21	1			
	6	3					1	14			3
	7								23		
	8							3		19	
	9	2						1		1	19

Fig. 3. HMM classifier confusion matrix for subject S1

		HMM Word Classification									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Word Spoken	**0**	21							1		
	1		18		3				2		
	2			22							
	3				21	1	1				
	4				1	21					
	5						22		1		
	6	1						19		1	1
	7								23		
	8	1		2	1		1	4		12	1
	9	1			2			1		6	12

Fig. 4. HMM classifier confusion matrix for subject S2

the vocabulary could account for the majority of the misclassifications. These results suggest that the HMM classifier would perform well in a multi-modal ASR system.

To test this hypothesis a separate identification multi-modal ASR system would have to be implemented, consisting of two uni-modal classifiers or experts: a MES ASR expert and an acoustic ASR expert. A supervisor could then fuse the output from both experts to decide the final classification. We anticipate that the acoustic expert would be able to accurately classify words that the MES expert has difficulties with, and that the MES expert would be able to accurately classify words that the acoustic expert has difficulties with. In addition to normal conditions, the performance of the multi-modal ASR system should also be evaluated under various stress conditions typical during flight of a jet aircraft (e.g. acoustic noise, positive pressure breathing)

V. Conclusions

It has been demonstrated that MES ASR using a HMM classifier is resilient to temporal variance, which offers improved robustness compared to the LDA classifier. The overall performance of the MES ASR can be further enhanced by optimizing the features and structure of the HMM classifier to improve classification rate. Nevertheless, the HMM classifier has already shown that it would effectively complement an acoustic classifier in a multi-modal ASR system.

Acknowledgment

This work was supported by the Defense Evaluation Research Agency, United Kingdom and the National Sciences and Engineering Research Council of Canada.

References

[1] Research and Technology Organization (North Atlantic Treaty Organization), "Alternative control technologies," RTO Technical Report 7, 1998.

[2] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," IEEE Trans. Signal Processing, vol. 40, no. 4, pp. 725-735, 1992.

[3] S. E. Bou-Ghazale, J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Trans. Speech Audio Processing, vol. 8, no. 4., pp. 429-442, 2000.

[4] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Myoelectric signals to augment speech recognition," Med. Biol. Eng. Comput., in press.

[5] M. Zardoshti, B. C. Wheeler, K. Badie, R. Hashemi, "Evaluation of EMG features for movement control of prostheses," Proceedings of the 15th Annual International Conference of the IEEE EMBS, vol. 15, no. 3, pp. 1141-1142, 1993.

[6] L. Rabiner, B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall: Englewood Cliffs, NJ, 1993.